The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications

The codon adaptation index - a measure of directional synonymous codon usage bias, and its potential applications

Paul M.Sharp[1,2*] and Wen-Hsiung Li[2]

[1]Department of Genetics, Trinity College, Dublin 2, Ireland and [2]Center for Demographic and Population Genetics, University of Texas, PO Box 20334, Houston, TX 77225, USA

Received August 4, 1986; Revised and Accepted December 18, 1986

ABSTRACT

A simple, effective measure of synonymous codon usage bias, the Codon Adaptation Index, is detailed. The index uses a reference set of highly expressed genes from a species to assess the relative merits of each codon, and a score for a gene is calculated from the frequency of use of all codons in that gene. The index assesses the extent to which selection has been effective in moulding the pattern of codon usage. In that respect it is useful for predicting the level of expression of a gene, for assessing the adaptation of viral genes to their hosts, and for making comparisons of codon usage in different organisms. The index may also give an approximate indication of the likely success of heterologous gene expression.

INTRODUCTION

The determination of the DNA sequences of a large number of genes from a wide variety of species has revealed that, in a large proportion of cases, the alternative synonymous codons for any one amino acid are not used randomly (1, and references therein). Further, it has been noted that a part of this nonrandom usage is species, or rather taxon, specific (2). However, within species there is considerable heterogeneity between genes, and in the two best studied organisms, namely Escherichia coli and the yeast Saccharomyces cerevisiae, there is a clear positive correlation between degree of codon bias and level of gene expression (3,4). Examination of large data sets from these species reveals that within species differences are largely in the degree rather than the direction of codon usage bias (5,6).

For many reasons it is desirable to quantify the degree of bias in codon usage in each gene in such a way that comparisons can be made both within and between species. One approach to this problem is to devise a measure for assessing the degree of deviation from a postulated impartial pattern of usage. The codon preference bias proposed by McLachlan et al. (7) is such a measure. Recently Sharp et al. (5) have proposed to calculate the chi square value for the deviation from random codon usage and then scale

the value by the gene length (number of codons) so that comparisons can be made between genes.

Another approach is to assess the relative merits of different codons from the viewpoint of translational efficiency. For example, Ikemura (1,8,9) has identified certain "optimal" codons in E.coli and yeast which are expected to be translated more efficiently than others, and calculated the frequency of optimal codons in a gene. The codon bias index of Bennetzen and Hall (4), for use with yeast genes, is essentially similar. Such indices are certainly useful, but have several disadvantages. First, some amino acids are usually excluded because it is not clear which codons are "optimal". Second, all codons considered are classified into only two categories, i.e., optimal and nonoptimal, with no recognition that some codons within each category are better than others. Third, there is no good basis for comparison between species because the proportional division of the codon table into the two categories may differ; e.g., Ikemura (1) identified 21 optimal codons for 14 amino acids in E.coli, and 19 optimal codons for 13 amino acids in yeast.

Gribskov et al. (10) have recently proposed another index, the codon preference statistic. This statistic is based on the ratio of the likelihood of finding a particular codon in a highly expressed gene to the likelihood of finding that codon in a random sequence with the same base composition as that in the sequence under study. They show that the statistic is useful for locating genes in sequenced DNA, for predicting the relative level of their expression, and for detecting sequencing errors. However, the statistic is not normalized and therefore the values for two genes encoding proteins with different amino acid compositions can be quite different even if both genes use only the "best" codons.

With various purposes in mind we have devised a new index. It is similar to the codon preference statistic but is normalized so that it is convenient for making comparisons both within and between species. After describing the index, we show some rather varied applications and indicate certain advantages over other indices. In recognition of the role of natural selection in producing high levels of codon bias, we call this statistic the Codon Adaptation Index.

METHODS

We recognize that even in E.coli and yeast the factors determining the frequency of synonymous codon usage are not completely understood, but that

several points are clear: the pattern of codon usage in any particular gene is largely determined by natural selection and mutation (5,6); selection appears to occur via translational efficiency, so that synonymous codon usage in highly expressed genes is under the strongest selective constraints (4,8,9); in E.coli and yeast, very highly expressed genes appear to have the greatest degree of synonymous codon bias (3-6,8). From these points it is deduced that the pattern of codon usage in very highly expressed genes can reveal (i) which of the alternative synonymous codons for an amino acid is the most efficient for translation, and (ii) the relative extent to which other codons are disavantageous.

The first step is, then, to construct a reference table of relative synonymous codon usage (RSCU) values from very highly expressed genes of the organism in question. An RSCU value for a codon is simply the observed frequency of that codon divided by the frequency expected under the assumption of equal usage of the synonymous codons for an amino acid (5). Thus,

$$RSCU_{ij} = \frac{X_{ij}}{\frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij}} \qquad [1]$$

where X_{ij} is the number of occurrences of the jth codon for the ith amino acid, and n_i is the number (from one to six) of alternative codons for the ith amino acid. The relative adaptiveness of a codon, w_{ij}, is then the frequency of use of that codon compared to the frequency of the optimal codon for that amino acid:

$$w_{ij} = RSCU_{ij} / RSCU_{imax} = X_{ij}/X_{imax} \qquad [2]$$

where $RSCU_{imax}$ and X_{imax} are the RSCU and X values for the most frequently used codon for the ith amino acid.

Codon usage data have been compiled previously for 165 genes from E.coli (6), and for 110 genes from yeast (5). To obtain reference RSCU values, we have taken the 27 very highly expressed E.coli genes compiled by Sharp and Li (6), which include genes encoding 17 ribosomal proteins, four outer membrane proteins and four elongation factors. For yeast a set of 24 genes has been taken from the high expression group previously identified (5). These include 16 genes encoding ribosomal proteins, one for an elongation factor, and seven loci encoding very abundant enzymes. The RSCU

Table 1. Values of RSCU and w for codons in very highly expressed genes from E.coli and yeast.

		E.coli		Yeast				E.coli		Yeast	
		RSCU	w	RSCU	w			RSCU	w	RSCU	w
Phe	UUU	0.456	0.296	0.203	0.113	Ser	UCU	2.571	1.000	3.359	1.000
	UUC	1.544	1.000	1.797	1.000		UCC	1.912	0.744	2.327	0.693
Leu	UUA	0.106	0.020	0.601	0.117		UCA	0.198	0.077	0.122	0.036
	UUG	0.106	0.020	5.141	1.000		UCG	0.044	0.017	0.017	0.005
Leu	CUU	0.225	0.042	0.029	0.006	Pro	CCU	0.231	0.070	0.179	0.047
	CUC	0.198	0.037	0.014	0.003		CCC	0.038	0.012	0.036	0.009
	CUA	0.040	0.007	0.200	0.039		CCA	0.442	0.135	3.776	1.000
	CUG	5.326	1.000	0.014	0.003		CCG	3.288	1.000	0.009	0.002
Ile	AUU	0.466	0.185	1.352	0.823	Thr	ACU	1.804	0.965	1.899	0.921
	AUC	2.525	1.000	1.643	1.000		ACC	1.870	1.000	2.063	1.000
	AUA	0.008	0.003	0.005	0.003		ACA	0.141	0.076	0.025	0.012
Met	AUG	1.000	1.000	1.000	1.000		ACG	0.185	0.099	0.013	0.006
Val	GUU	2.244	1.000	2.161	1.000	Ala	GCU	1.877	1.000	3.005	1.000
	GUC	0.148	0.066	1.796	0.831		GCC	0.228	0.122	0.948	0.316
	GUA	1.111	0.495	0.004	0.002		GCA	1.099	0.586	0.044	0.015
	GUG	0.496	0.221	0.039	0.018		GCG	0.796	0.424	0.004	0.001
Tyr	UAU	0.386	0.239	0.132	0.071	Cys	UGU	0.667	0.500	1.857	1.000
	UAC	1.614	1.000	1.868	1.000		UGC	1.333	1.000	0.143	0.077
ter	UAA	--	--	--	--	ter	UGA	--	--	--	--
ter	UAG	--	--	--	--	Trp	UGG	1.000	1.000	1.000	1.000
His	CAU	0.451	0.291	0.394	0.245	Arg	CGU	4.380	1.000	0.718	0.137
	CAC	1.549	1.000	1.606	1.000		CGC	1.561	0.356	0.008	0.002
Gln	CAA	0.220	0.124	1.987	1.000		CGA	0.017	0.004	0.008	0.002
	CAG	1.780	1.000	0.013	0.007		CGG	0.017	0.004	0.008	0.002
Asn	AAU	0.097	0.051	0.100	0.053	Ser	AGU	0.220	0.085	0.070	0.021
	AAC	1.903	1.000	1.900	1.000		AGC	1.055	0.410	0.105	0.031
Lys	AAA	1.596	1.000	0.237	0.135	Arg	AGA	0.017	0.004	5.241	1.000
	AAG	0.404	0.253	1.763	1.000		AGG	0.008	0.002	0.017	0.003
Asp	GAU	0.605	0.434	0.713	0.554	Gly	GGU	2.283	1.000	3.898	1.000
	GAC	1.395	1.000	1.287	1.000		GGC	1.652	0.724	0.077	0.020
Glu	GAA	1.589	1.000	1.968	1.000		GGA	0.022	0.010	0.009	0.002
	GAG	0.411	0.259	0.032	0.016		GGG	0.043	0.019	0.017	0.004

Genes used:
E.coli - 17 ribosomal protein genes, 4 elongation factor genes, 4 outer membrane protein genes, recA, dnaK (data from Ref.6)
Yeast - 16 ribosomal protein genes, TEF 1, 2 enolase genes, 2 GA-3-PDH genes, ADH 1, PGK, pyruvate kinase (data sources given in Ref.5)

and w values obtained for very highly expressed genes from E.coli and yeast are given in Table 1.

The Codon Adaptation Index (CAI) for a gene is then calculated as the geometric mean of the RSCU values (from Table 1) corresponding to each of the codons used in that gene, divided by the maximum possible CAI for a gene of the same amino acid composition, i.e.,

$$CAI = CAI_{obs} / CAI_{max} \quad [3]$$

where

$$CAI_{obs} = \left(\prod_{k=1}^{L} RSCU_k \right)^{1/L} \quad [4]$$

$$CAI_{max} = \left(\prod_{k=1}^{L} RSCU_{kmax} \right)^{1/L} \quad [5]$$

where $RSCU_k$ is the RSCU value for the kth codon in the gene, $RSCU_{kmax}$ is the maximum RSCU value for the amino acid encoded by the kth codon in the gene, and L is the number of codons in the gene.

Note that if a certain codon is never used in the reference set then the CAI for any other gene in which that codon appears becomes zero. To overcome this problem we assign a value of 0.5 to any X_{ij} that would otherwise be zero. Also, the number of AUG and UGG codons are subtracted from L, since the RSCU values for AUG and UGG are both fixed at 1.0, and so do not contribute to the CAI.

As illustration, consider the rpsU gene from E.coli which, excluding the initiation codon, comprises 70 codons and has the sequence:

.CCG.GTA.ATT.AAA.GTA. . . . .

For that sequence and from the RSCU values in Table 1:

$$CAI_{obs} = (3.288 \times 1.111 \times 0.466 \times 1.596 \times 1.111 \times .....)^{1/70}$$

and $$CAI_{max} = (3.288 \times 2.244 \times 2.525 \times 1.596 \times 2.244 \times .....)^{1/70}$$

From these two values and equation [3] we can obtain the CAI value.
We note that equation [3] is exactly equivalent to:

$$CAI = \left(\prod_{k=1}^{L} w_k \right)^{1/L} \quad [6]$$

Table 2. CAI values for E.coli and yeast genes.

E.coli gene	CAI	yeast gene	CAI
17 RPs	0.467-0.813	16 RPs	0.529-0.915
rpsU	0.726	histones	0.532-0.733
rpoD	0.582		
dnaG	0.271	2u plasmid	0.099-0.106
lacI	0.296	GAL 4	0.116
trpR	0.267	PPR 1	0.114
lpp	0.849[a]	GPD 1	0.929[a]
hsdS	0.218[b]	mat A2	0.098[b]

RPs - ribosomal protein genes.
a highest CAI value among data set.
b lowest CAI value among data set.

where w_k is the w value for the kth codon in the gene (see equation [2]). Therefore, for rpsU:

$$CAI = (1.00 \times 0.495 \times 0.185 \times 1.000 \times 0.495 \times .....)^{1/70}$$

Equation [6] saves computation time. To overcome real number underflow problems in computer calculations, equation [6] can be computed as:

$$CAI = \exp \frac{1}{L} \sum_{k=1}^{L} \ln w_k \qquad [7]$$

or from a codon usage table:

$$CAI = \exp \frac{1}{L} \sum_{i=1}^{18} \sum_{j=1}^{n_i} X_{ij} \ln w_{ij} \qquad [8]$$

where X_{ij} and n_i are as defined in equation [1].

There is no intrinsic effect of gene length (L) on CAI, but CAI values from short genes may be more variable due to sampling effects.

APPLICATIONS and DISCUSSION

Predicting levels of gene expression within a species.

CAI values clearly parallel levels of gene expression. Ribosomal protein genes are highly expressed, and have generally high CAI values

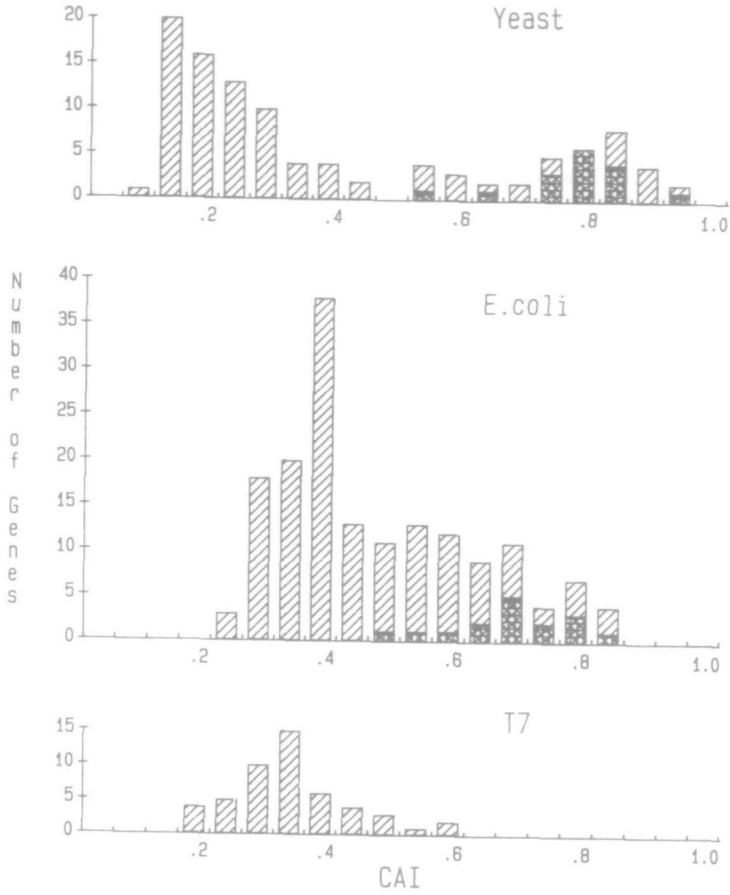

Figure 1. Distribution of CAI values for (a) 106 yeast genes, (b) 165 E.coli genes, and (c) 50 bacteriophage T7 genes. In (a) and (b) ribosomal protein genes are cross-hatched. Plasmid genes are excluded.

(Table 2, Figure 1). Among yeast ribosomal protein genes only that encoding S33 has a CAI < 0.6, and it is a very short gene (L = 65). Lowly expressed regulatory genes (e.g., lacI, trpR in E.coli; GAL 4, PPR 1 in yeast) have low CAI values (Table 2). In E.coli the relationship between codon bias and gene expression is perhaps best illustrated by considering operons (as suggested by Gouy and Gautier, Ref.3). For example, within the macromolecular synthesis operon the expression levels are rpsU >> rpoD >> dnaG (11), and the CAI values for these genes are 0.726, 0.582 and 0.271, respectively (Table 2). Eight of the nine genes of the unc operon encode the

Table 3. CAI values for genes in the unc operon of E.coli.

Pos	Gene	CAI	L	Gene Product name	amount		sector
1	papI	0.238	127		??		
2	papD	0.400	253	chi	1	:	
3	papH	0.583	71	omega	10	:	F_0
4	papF	0.482	152	psi	2	:	
5	papE	0.374	169	delta	1	:	
6	papA	0.665	501	alpha	3	:	
7	papC	0.403	273	gamma	1	:	F_1
8	papB	0.650	444	beta	3	:	
9	papG	0.474	133	epsilon	1	:	

Pos : gene position within the operon (1 - 5').
The relative amount of each gene product in the ATPase complex is taken from Ref.12.

eight subunits of the F_0 and F_1 sectors of the H^+-ATPase complex, and the stoichiometry of these subunits is known (12). The CAI value is clearly correlated with the level of gene expression among the genes encoding subunits of the F_1 sector (Table 3), with the CAI values for papA and papB being similar, and much higher than those for papE, papC and papG. Among genes encoding subunits in the F_0 sector the rank order of CAI values corresponds to the relative amounts of the gene products required. The CAI for papH is perhaps surprisingly low, but this is a very short gene (Table 3). The function of papI is unknown. The CAI value for papI is very low, and may indicate that this is a regulatory gene, or perhaps (see below) a noncoding open reading frame.

Although many of the measures of codon bias discussed in the Introduction seem to be positively correlated with gene expression, we feel that CAI has the twin advantages of being simple to calculate and making greater quantitative use of available information (see 'Comparison of CAI with other indices' below).

The positive correlation between degree of synonymous codon bias and expression level in E.coli (and yeast) seems firmly established, but the causal relationship between the two has been debated. We have concluded elsewhere (6) that the degree of codon bias reflects the past action of natural selection -- it is indicative of the level at which the gene is expressed, rather than dictating that level. This seems to concur with conclusions drawn from a theoretical model of the translation process (13).

Table 4. CAI values for mammalian genes using E.coli and yeast RSCU values.

Heterologous gene	Host	
	E.coli	Yeast
Human alpha interferon	0.218	0.099
Human insulin	0.307	0.043
Human growth hormone	0.287	0.082
Human factor VIII	0.205	0.114
Human factor IX	0.263	0.176
Bovine chymosin	0.326	0.086

Predicting levels of heterologous gene expression.

There is experimental evidence that certain codons can affect expression level (14-17). For example, the AGG codon markedly affects the translation rate of genes in E.coli (14,15). This suggests that for a heterologous gene to have a maximal level of expression its codon usage must correspond to that of the host. By using the RSCU values of potential hosts to calculate CAI values for a heterologous gene it should be possible to predict how well suited that gene would be to the translational systems of those hosts. In Table 4 the CAI values of some genes of biotechnological interest are given for two different potential hosts, E.coli and yeast. In each case these mammalian genes seem better 'adapted' to E.coli, suggesting that high expression might be more easily obtained in that system. Of course, in reality, the choice of host would probably depend on other practicalities. The CAI would, however, suggest whether it is likely to be either necessary or of any benefit to chemically synthesize a new gene, to include more appropriate codons. It should be stressed that the CAI is only an approximate indication of the suitability of the codon usage within a gene. For example, it takes no account of the distribution of codons along the gene, yet theoretical considerations suggest that this may be very important (18).

A measure of evolutionary adaptedness.

Under certain natural circumstances foreign genes are expressed in host organisms. Viral genes are an obvious example. Codon usage in the many bacteriophages which do not encode their own tRNA molecules should be adapted to the translational machinery of the host. Then the CAI, using host RSCU values, is an estimate of the degree of adaptation. For example, comparison of the pattern of codon usage in the genes of bacteriophage T7

Table 5. CAI values for homologous genes from E.coli and T7.

E.coli gene	CAI	T7 gene	CAI
ssb	0.605	2.5	0.573
dnaG	0.271	4	0.301
polA	0.391	5	0.341
		6	0.387

with the relative abundance of cognate tRNA molecules in E.coli (considered to be the usual host of T7) suggests that T7 genes are not so well adapted as E.coli's own genes, although there is clearly some adaptation (19,20). This seems to be confirmed by contrasting the distribution of CAI values for T7 genes with those of E.coli (Figure 1). However, the difference seen in Figure 1 could arise in part because the genes contrasted encode different products; for example, T7 encodes no ribosomal proteins. It has been reported that four genes in T7 are homologous to three E.coli genes (21). A comparison of these genes (Table 5) is not conclusive, because only ssb is highly adapted in E.coli, although in that case the T7 gene does have a lower CAI. The four T7 genes as a group do not seem to be significantly less adapted than the three E.coli genes.

In cases where it has not been clear which organism represents the major host for a virus it may prove informative to calculate CAI values with the different RSCU values of potential hosts. For example, despite approximately 65% DNA homology between ØX174 and G4, the genomes of these two "coliphages" show a remarkable difference with respect to the frequency of the recognition sites of enterobacterial restriction enzymes (22). While ØX174 (as well as several other coliphages) has a significant avoidance of these sites, presumably reflecting adaptation to infecting E.coli, G4 does not. However, CAI values for the 10 genes of ØX174 and G4 are very similar, suggesting that the patterns of codon usage of the two phages are adapted (to E.coli) to equivalent extents.

Natural foreign gene expression would also occur if genes undergo horizontal transfer. Felmlee et al. (23) have discussed a possible example. They reported the DNA sequence of a region of the E.coli chromosome encoding four hemolysin genes, and found that their base composition and codon usage are atypical of that species. This, together with the observation that these genes are found in only a limited number of E.coli strains, was taken as

evidence that the genes represent a recent acquisition to this species (23). The CAI values for these genes are indeed very low, ranging from 0.202 to 0.243. These values are lower than those for nearly all other E.coli genes (see Figure 1, in which the hemolysin genes are not included), including some (e.g., araC and dnaG) which are expressed at very low levels. Hemolysin is an extracellular protein and would be expected to be expressed at much higher levels than araC or dnaG, so these low CAI values suggest that the hemolysin genes are not well adapted to E.coli, and seem to confirm the suggestion of a recent acquisition. If reference RSCU data were available for a variety of organisms from which the genes could have been transferred, it might be possible to determine the most likely source by comparison of CAI values.

If plasmids were regularly subject to interspecific transfer, then their genes might not become adapted to any one host. Genes on E.coli plasmids tend to have less codon bias than chromosomal genes (3). We note that the three genes of the yeast 2 micron plasmid have very low CAI values (Table 2).

Synonymous codon usage and the rate of molecular evolution.

A major prediction of the neutral theory of molecular evolution (24) is an inverse relationship between the rate of evolution and the degree of selective constraint, i.e., the stronger the constraint the slower the rate of molecular evolution. Indeed, a great deal of evidence confirms this, including the observation that pseudogenes, which are under no apparent constraint, are the fastest evolving DNA sequences (25). That synonymous substitutions in protein coding genes occur at a slower rate than substitutions in pseudogenes (26,27) implies that there are selective constraints on the former. If the differences between genes in degree of codon usage bias largely reflect differences in selection pressure on synonymous codons, then the rate of synonymous substitution would be inversely related to the degree of codon bias. The CAI can be used to quantify this relationship. Comparisons of E.coli and Salmonella typhimurium genes do indeed show a significant negative correlation between the rate of synonymous substitution and the CAI (28).

Comparison of codon usage in different organisms.

Meaningful comparisons of codon usage in different organisms can be made if care is taken in defining the reference set of genes from which the RSCU values are calculated. The reference sets we have chosen for E.coli and yeast comprise very similar collections of genes, yet the distribution of

CAI values for genes from these two organisms are rather different. Very highly expressed genes in yeast have on average a more extreme codon bias than their counterparts in E.coli, as seen for example with ribosomal protein genes (Table 2). The reference set of RSCU values reflects this, and so the genes with least codon usage bias in yeast have lower CAI values than genes in E.coli, as a result. It is particularly interesting to note that cluster analysis of yeast genes based on their synonymous codon usage clearly differentiates two groups, identified as comprising highly and moderately/lowly expressed genes (5), and that those two groups correspond almost exactly to the bimodal distribution of CAI values for yeast genes in Figure 1. By contrast, cluster analysis does not so easily differentiate highly and lowly expressed genes in E.coli or in T7 (5) and the distributions of CAI values from those organisms are unimodal (Figure 1). It is not clear why selection has apparently been more successful in producing high codon bias in yeast than in E.coli. Li (29) has shown that the effectiveness of selection in maintaining synonymous codon bias depends largely on the strength of selection and effective population size. It could be that the strength of selection is stronger in yeast than in E.coli because the required amount of certain gene products, such as ribosomal proteins, is larger. It is also possible that the effective population size is larger in yeast than in E.coli because the latter has a largely clonal population structure (30).

We note that comparisons between species can be difficult when the reference sets of genes have quite different levels of bias in codon usage. For example, very highly expressed genes have a much lower bias in codon usage in Bacillus subtilis than in E.coli or yeast (Shields and Sharp, in prep.). Then, in B.subtilis, there are few codons with very low w values. As a consequence, CAI values for other genes in B.subtilus are, on average, higher than those seen in the other species, even though the B.subtilus genes have clearly less bias. The CAI_{obs} given by equation [4] is less affected by this difference in the reference set, and may form a better basis for comparison between species under these circumstances.

Identification of protein-coding reading frames.

Several of the indices of codon usage bias were originally devised in order to ascertain the likelihood that open reading frames are indeed protein-coding. As with the other measures, the CAI should be useful in this context, particularly in locating genes of moderate to high expression. However, some of the points outlined above indicate that difficulties may

arise in interpreting low CAI values. Thus, while a high CAI is probably a good indication that a reading frame is protein-coding, a low CAI may indicate a gene of low expression, a gene of heterologous origin (as with the hemolysin genes), or a noncoding region that happens to contain no termination codons. The CAI value expected for a random sequence can easily be calculated, but a relatively high value for a noncoding sequence may arise simply because DNA is not a random sequence of nucleotides, or because there is a coding sequence on the complementary strand (31). For example, an E.coli gene with no UUA, CUA or UCA codons, but otherwise having the typical codon composition of a nonhighly expressed gene (6), would give rise to an in phase open reading frame on the complementary strand with a CAI of approximately 0.28, which is similar to the lower values seen for E.coli genes (Figure 1) and somewhat higher than the value (about 0.17) expected for a random sequence.

Comparison of CAI with other indices.

The CAI is a very simple measure of the extent of synonymous codon usage bias, specifically in the direction of the bias seen in highly expressed genes. It has the advantage, compared with indices which measure only the frequency of certain optimal codons, of taking account of all 59 codons where synonymous alternatives exist, each in a quantitative manner. For example, both the codon bias index (4) and the frequency of optimal codons (1) treat GCU and GCC equally, as preferred codons for Ala in yeast, and yet the frequency of GCU is approximately three times that of GCC in very highly expressed genes (Table 1). With heterologous gene expression in mind it may be of primary importance to know the frequency of particularly disadvantageous codons in a gene. Simpler indices compound these very rare codons with others not in the 'optimal' category. Thus in E.coli AUA and AUU are treated equally (1), despite their very different frequency of use (see Table 1, and Ref.6). Again the CAI takes account of these differences quantitatively.

The codon preference statistic (10) is similar but not identical to the CAI_{obs} given by equation [4]. One difference is that in calculating the codon preference statistic the p values (analagous to RSCU in equation [4]) are adjusted to take account of base composition. Another difference is that the CAI value is scaled to allow for the different amino acid compositions of different proteins (see equation [3]), and has a range from 0 - 1.0. Although this scaling cannot completely compensate for differing amino acid compositions, it facilitates comparisons between genes.

Our discussion of the use of the Codon Adaptation Index has focussed on unicellular organisms because the determinants of codon usage in multicellular organisms are not well understood (1). For example, it appears that the mammalian genome comprises regions of quite different G+C content (32), and that local G+C content is an important influence on codon usage in any one gene (1). Also tRNA abundancies are important selective constraints on codon usage, and in multicellular organisms tRNA populations vary among tissues. We also note that the only mammalian ribosomal protein genes for which DNA sequence data are available (two from mouse and two from rat -- see Ref.33) do not seem to show particularly high synonymous codon bias. It may be possible in the near future to derive a reference set of RSCU values from other highly expressed mammalian genes, and/or it may prove necessary to take into account the tissue in which the gene is expressed, for example by having several reference sets.

ACKNOWLEDGMENTS

We thank R. Grantham for suggestions and Robert J. Schwartz for plotting the figure. This study was supported by NIH grant GM 30998.

*To whom correspondence should be addressed

REFERENCES

1. Ikemura, T. (1985) Mol. Biol. Evol. 2, 13-34.
2. Grantham, R., Gautier, C., Gouy, M., Jacobzone, M. and Mercier, R. (1981) Nucleic Acids Res. 9, r43-r79.
3. Gouy, M. and Gautier, C. (1982) Nucleic Acids Res. 10, 7055-7074.
4. Bennetzen, J.L. and Hall, B.D. (1982) J. Biol. Chem. 257, 3026-3031.
5. Sharp, P.M., Tuohy, T.M.F. and Mosurski, K.R. (1986) Nucleic Acids Res. 14, 5125-5143.
6. Sharp, P.M. and Li, W-H. (1986) Nucleic Acids Res. 14, 7737-7749.
7. McLachlan, A.D., Staden, R. and Boswell, D.R. (1984) Nucleic Acids Res. 12, 9567-9575.
8. Ikemura, T. (1981) J. Mol. Biol. 146, 1-21.
9. Ikemura, T. (1982) J. Mol. Biol. 158, 573-598.
10. Gribskov, M., Devereux, J. and Burgess, R.R. (1984) Nucleic Acids Res. 12, 539-549.
11. Burton, Z.F., Gross, C.A., Watanabe, K.K. and Burgess, R.R. (1983) Cell 32, 335-349.
12. Foster, D.L. and Fillingame, R.H. (1982) J. Biol. Chem. 257, 2009-2013.
13. Holm, L. (1986) Nucleic Acids Res. 14, 3075-3087.
14. Robinson, M., Lilley, R., Ltttle, S., Emtage, J.S., Yarranton, G., Stephens, P., Millican, A., Eaton, M. and Humphreys, G. (1984) Nucleic Acids Res. 12, 6663-6671.
15. Bonekamp, F., Andersen, H.D., Christensen, T. and Jensen, K.F. (1985) Nucleic Acids Res. 13, 4113-4123.
16. Pedersen, S. (1984) EMBO J. 3, 2895-2898.
17. Varenne, S., Buc, J., Lloubes, R. and Lazdunski, C. (1984) J. Mol. Biol. 180, 549-576.

18. Varenne, S. and Lazdunski, C. (1986) J. Theor. Biol. 120, 99-110.
19. Sharp, P.M., Rogers, M.S. and McConnell, D.J. (1985) J. Mol. Evol. 21, 150-160.
20. Grantham, R., Greenland, T., Louail, S., Mouchiroud, D., Prato, J.L., Gouy, M. and Gautier, C. (1985) Bull. Inst. Pasteur 83, 95-148.
21. Toh, H. (1986) FEBS Letters 194, 245-248.
22. Sharp, P.M. (1986) Mol. Biol. Evol. 3, 75-83.
23. Felmlee, T., Pellett, S. and Welch, R.A. (1985) J. Bact. 163, 94-105.
24. Kimura, M. (1983) The Neutral Theory of Molecular Evolution. Cambridge University Press, Cambridge.
25. Li, W-H., Luo, C-C and Wu, C-I. (1985) in Molecular Evolutionary Genetics, MacIntyre, R.J. Ed., pp. 1-94, Plenum Press, New York.
26. Li, W-H., Gojobori, T. and Nei, M. (1981) Nature 292, 237-239.
27. Miyata, T. and Hayashida, H. (1981) Proc. Natl. Acad. Sci. USA 78, 5739-5743.
28. Sharp, P.M. and Li, W-H. (1986) J. Mol. Evol. (in press)
29. Li, W-H. (1987) J. Mol. Evol. (in press)
30. Ochman, H. and Selander, R.K. (1984) Proc. Natl. Acad. Sci. USA 81, 198-201.
31. Sharp, P.M. (1985) Nucleic Acids Res. 13, 1389-1397.
32. Bernardi, G., Olofsson, B., Filipski, J., Zerial, M., Salinas, J., Cuny, G., Meunier-Rotival, M. and Rodier, F. (1985) Science 228, 953-958.
33. GenBank, Genetic Sequence Data Bank, Release 42.0 (1986) Bolt, Beranek and Newman.